How Should We Measure Individual Researcher‘s Performance Capacity Within and Between Universities? A Multilevel Extension of the Bibliometric Quotient (BQ)

Mutz, Rüdiger ; Daniel, Hans-Dieter

Originally published at:
Mutz, Rüdiger; Daniel, Hans-Dieter (2019). How Should We Measure Individual Researcher‘s Performance Capacity Within and Between Universities? A Multilevel Extension of the Bibliometric Quotient (BQ). In: Catalano, G; Daraio, C; Gregori„ M; Moed„ H F; Ruoca, G. Proceedings of the 17th International Conference on Scientometrics Informetrics ISSI, September 2-5, 2019 Sapienza University Rom, Italy. Rom, Italy: International Society for Scientometrics and Informetrics, 1098-1109.

ISSI 2019

SAPIENZA
UNIVERSITÀ DI ROMA

17th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

ISSI2019

with a Special STI Indicators Conference Track

2-5 September 2019
Sapienza University of Rome, Italy

PROCEEDINGS

VOLUME I

EdizioniEfesto

How Should We Measure Individual Researcher`s Performance Capacity Within and Between Universities – Social Sciences as an Example? A Multilevel Extension of the Bibliometric Quotient (BQ)

Rüdiger Mutz[1] and Hans-Dieter Daniel[2, 3]

[1] *mutz@gess.ethz.ch*
ETH Zurich, Professorship for Social Psychology and Research on Higher Education, Andreasstrasse 15, CH-8050 Zurich (Switzerland)

[2] *daniel@gess.ethz.ch*
ETH Zurich, Professorship for Social Psychology and Research on Higher Education, Andreasstrasse 15, CH-8050 Zurich (Switzerland)

[3] *hans-dieter.daniel@uzh.ch*
University of Zurich, Department of Psychology, Binzmuehlestrasse 14, CH-8050 Zurich (Switzerland)

Abstract

The assessment of individual research performance has become a major attraction for bibliometric researchers in recent years, and is dominated by the classic bibliometric indicator approach (e.g., h-index). Alternatively, a psychometric measurement approach is favored, which considers measurement errors. It is assumed that the "researcher's performance capacity" as a personal trait and competency is responsible for the individual research performance, which might vary randomly due to measurement errors. Five individual-level bibliometric variables served as items (e.g., number of articles in top 5%) to measure the competency. The central question of this contribution is how much variance in the "researcher`s performance capacity" is explained by differences between universities/subfields. With bibliometric data (Scopus) for a sample of 1,071 social scientists with Swiss university affiliations a one-dimensional scale ("Bibliometric Quotient", BQ) was created by means of a psychometric model, which has a high, but not perfect, reliability of r_{tt}=.84. The items were most suitable for scientists scoring above average. About 33% of the variance of the BQ is due to differences between the universities/subfields, and only 7% of the variance is due to differences between universities alone. A ranking only of Swiss universities in the social sciences does not necessarily make sense.

Introduction

The bibliometric-based measurement of individual research performance has attracted a great deal of attention in recent years, which is reflected in a multitude of literature on this topic (e.g., Abramo & D'Angelo, 2014; Bornmann & Marx, 2013, 2014; Bornmann & Mutz, 2011; Wildgaard, Schneider, & Larsen, 2014). "The evaluation of individual research performance is a fundamental tool for management, to inform decisions in areas such as faculty recruitment, career advancement, reward systems, grants awarding and projects funding." (Abramo, Cicero, & D`Angelo, 2013, p. 528). A large number of numerical indicators were developed at the level of the individuals. Wildgaard et al. (2014, p. 125) "reviewed 108 indicators that can potentially be used to measure performance on individual author-level". A prototype for such an indicator to assess individual research performance is the h-index.

The indicator approach, more or less adopted from economics, sociology and natural sciences, is less widespread in the sciences that deal with the individual, namely psychology or educational sciences. A major reason for this is the problem of random measurement errors, which are more significant at the level of individuals than at the level of institutions, and which are often not taken into account in the indicator approach (Abramo, D`Angelo, & Grilli, 2015; Karlsson et al., 2015). Due to different coverages of bibliographic databases, single publications

of individual researchers may be missing. Citation fluctuations might occur as a result of database updates (inclusion or removal of journals). Single highly cited publications do not reflect the overall work of a researcher. Such random fluctuations usually do not play a role at the institutional level, since they are averaged out during aggregation, especially if the size of institutions is high. Instead of relying on single indicators, psychology and educational sciences use a set of “indicators” called “items” that homogeneously measure a characteristic as theoretical construct that is not itself directly observable. These items have only a meaning within the construct they measure and may also be affected by measurement errors. A variety of psychometric test models have been developed to estimate quantitative test scores from empirical test data and thus measure a person's trait as a time-stable behavioral tendency. One characteristic, which has become generally known, is "intelligence": "A global concept that involves an individual's ability to act purposefully, think rationally, and deal effectively with the environment" (Wechsler, 1958, p. 7).

This measurement perspective motivated us to create a psychometric model based on bibliometric data to capture the scientific performance of researchers. With a modeling approach we hope to clarify questions of reliability, validity and fairness of the scale, and questions of dimensioning. These questions often remain unanswered in the classic indicator approach. A first model and a scale, the so-called Bibliometric Quotient, has already been developed and applied exemplarily to data from a sample of researchers in the field of social science methodology (Mutz & Daniel, 2018). Models have the advantage that they can be extended at will. Specific problems of a model can be solved by adding further model components in the hope the model fits the data better. In the indicator approach, special problems of an indicator (e.g., h-index) are often solved by the development of new indicators, whereby the letters of the alphabet are no longer sufficient to name the multitude of indicators (e.g., h-index, b-index, M-index), which has been developed.

This paper aims to extend the previous psychometric model of the Bibliometric Quotient (BQ) by a multilevel component, which considers differences between and within institutions of higher education. A topic, which attracts attention in the bibliometric indicator research, as well (Abramo, Cicero, & D`Angelo, 2012; Bonaccorsi & Cicero, 2016). How much variance in the BQ is explained by differences between universities? Institutional comparisons and rankings require a sufficient variability between institutions compared to the variability within institutions. The approach will be applied to bibliometric data on social scientists with Swiss university affiliations. The following research questions are in the focus:

1) Is it possible to create a one-dimensional scale form bibliometric data in order to measure the researcher's performance capacity? How reliable is the scale?
2) How high is BQ of social scientists from Switzerland?
3) How much variance in the BQ is explained by differences between universities? Is it possible to rank Swiss universities?
4) How strong are the relationships between the BQ and classic bibliometric indicators (e.g., h-index, total citations)?

Psychometric measurement model

Adopting the person-environment approach from psychology, we assume that the scientific performance of a researcher in the form of publications and their (citation) impact on the scientific field is based on the stable disposition or competency of a researcher (*person*), and on the research *environment*, in which he or she works (e.g., high citation level in life sciences). This competency is called “researcher`s performance capacity” (Harnad, 2008), as "competency of the researchers as authors to write influential papers" (Mutz & Daniel, 2018, p. 1284). To measure the theoretical construct, you need some individual specific variables,

called items, that repeatedly measure the same construct. In the case of measurement errors, it is expected that any kind of aggregation of items (i.e. scale) is more reliable than any single item. We assume that the "researcher's performance capacity" is the higher (in brackets the item labels),

- the higher the *scientific impact of the researcher's articles* in the researcher's scientific field is, measured as the number of publications that are in the top 5% in a scientific field (top5%, ITEM 1),
- the more that the publications have published as *first-author*, *mainly responsible for the article* (number of first author paper, ITEM 2),
- the higher the *impact of a single article*, the citation of the highest cited paper is (citation of the highest cited paper, ITEM 3),
- the more articles of a researcher have citations beyond the mean citation level of a field (MNCS, ITEM 4)
- the stronger the *short-term resonance* of the researcher in the scientific community is, measured as the total number of citations of the researcher's publications in a 5-year citation window (total citations in 5-year window, ITEM 5).

The *Item characteristic curve Poisson counts model* (ICCPCM) by Doebler, Doebler, and Holling (2014) serves as a psychometric model. It starts from a binary Rasch model as core model and add a frame model, which transforms the binary model to a Poisson count model.

Expressed in simple terms, the binary Rasch model, firstly introduced in bibliometrics by Alvarez and Pulgarin (1996a, 1996b, 1996c), assumes that the probability of an individual reaction to a binary item, for example, the probability that a researcher has published at least one paper in a relevant international journal or not, is a function of both the difficulty of the respective item, and the researcher`s competency (Andrich, 2010). For researchers with high competency (i.e., researcher's performance capacity), it is easier to publish in an international journal than for researchers with low competency. With increasing researcher's performance capacity, the probability of being able to publish in an international journal increases. This can be represented as an s-shaped exponential function, the so called *Item Characteristic Curve* (ICC), where the probability ranges from 0 (= no publication) to 1 (= publication), where 0 and 1 are only approximated and never reached (Fig 1.).

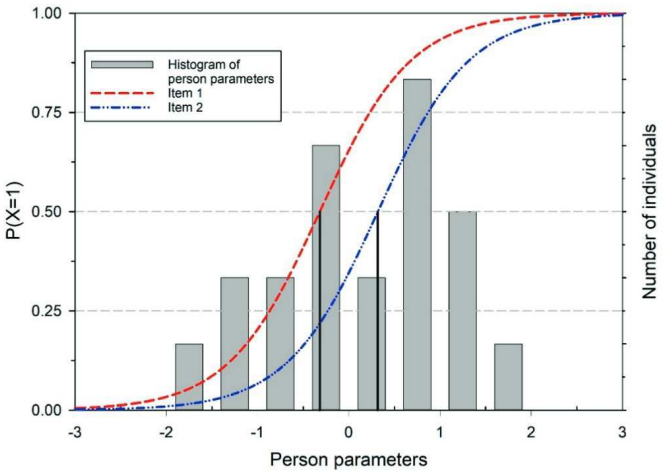

Fig. 1. Item characteristic curves for two items and histogram of person parameters (fictitious data).

However, the personal competency alone is not sufficient. The difficulty of the item must also be taken into account, as well, which represent the environment component. Thus it is much more difficult for a researcher to publish an article in a high-impact journal (e.g., “Nature”) (Item 2, Fig. 1) than in a low impact journal (Item 1, Fig. 1). Therefore, the ICC of Item 2 is shifted to the right on the x-axis in comparison to the ICC of Item 1.

In addition, items can separate differently well between researchers with high competency compared to researchers with low competency. For example, a single publication in a high-impact journal (e.g., "Nature") might already identify a researcher as excellent, since he or she then scores high in all other items as well. The item (e.g., to publish in "Nature" or not) would then have a high item discrimination as the second item parameter of the Rasch model. The s-shaped curve would be very steep (see Item 2, Fig.1).

Since bibliometric raw data (e.g., number of citations) are usually counts (i.e., integer numbers including zero), the binary Rasch model must be transformed to a count model. This is done by multiplying the binary core model with a g-component, which indicates the maximum expected count of an item in the sample. A problem is that researchers, who had more life time for their research (“active research time”) are favored over researchers with less life time, because they have more time to publish. For this reason, the active research time is also included in the model. In the last step, the model is extended to consider the impact of institutions by dividing the person parameter (histogram, Fig. 1) into two components: an institutional component and an individual-specific residual component within the institution.

Two assumptions of the model are of particular importance: Specific objectivity and local stochastic independency. According to the assumption of "specific objectivity", differences between items (e.g., item difficulties) should be independent of the sample of individuals, which were assessed, and vice versa, differences between individuals should be independent of which items are used to assess the individuals. A simple way to check this is to divide the data set into two groups (e.g., 2 subfields) and test whether the item parameters differ between the two groups. The local stochastic independency assumption assumes that the person parameters are the sole cause of the correlations among the items. If the relationships among them are statistically controlled for the person parameters, the resulting residuals are uncorrelated in the case of "local stochastic independency".

All model parameters including the person parameters can be estimated with the *two-parameter ICC Poisson counts model*, which can be formalized as follows: For $i = 1$ to N_i items as random variables X_{vi} with realized count outcomes x_{vi} for $v = 1$ to N_v researchers, the expected value in counts for the final Poisson-distributed random variable X_{vi} is (Mutz & Daniel, 2018):

$$\lambda_{vi} = E(X_{vi} \mid \varphi(g, \beta_i, \xi_v, \alpha_i)) = g \cdot time_v \left[\frac{e^{\alpha_i(\xi_v - \beta_i)}}{1 + e^{\alpha_i(\xi_v - \beta_i)}} \right] \quad (1)$$

$$X_{vi} \sim \text{Poisson}(\lambda_{vi}),$$

where g is the maximum annual value of X_{vi} (e.g., the maximal annual number of publications in the sample of individual researchers),
ξ_v is the person parameter of individual v,
β_i is the item parameter or item difficulty of item i,
α_i is the discrimination parameter for item i (the higher the value is, the more the item discriminate between individuals with high or low competency),

$rtime_v$ is the observed active research time of researcher *v* (the year of the last publication minus the year of the first publication of a researcher *v*) .

Due to the fact that the Poisson distribution is very restrictive (the mean value is equal to the variance), the Poisson distribution have often to be replaced with the Negative Binomial distribution (Mutz & Daniel, 2018; Mutz & Wolbring, 2017). In order to represent the variability between institutions, the person parameter is again divided into two components as follows (Fox, 2010, p. 145f):

$$\theta_v = \theta_{v(h)} + \theta_h, \quad (2)$$

where $\xi_{v(h)}$ represents the individual specific component within the institution h (residual) and γ_h the effect of institution h (2-level model). The model can be estimated by a Bayesian estimation approach suggested by Stone and Zhu (2015). The within variance of $\sigma^2_{\xi v(h)}$ is fixed to 1.0 (informative prior).

Data and Methods

The publisher Elsevier provided us with bibliometric raw data from the bibliographic database Scopus to ~500,000 publications from all subject areas published between 1996 and 2015, in which at least one author with a Swiss university affiliation was involved. A comprehensive data cleansing was carried out, which mainly concerned the affiliations. The publications often use different spellings from the same institution (e.g., EPF Lausanne, EPFL, ETHL, Swiss Federal Institute Lausanne, Swiss Federal Institute of Technology Lausanne, École polytechnique féderale de Lausanne), some of which Scopus provided with a different organization ID.

Since the analysis does not primarily refer to publications, but to researchers, a sample was drawn from social scientists with the following characteristics: *experienced researchers from the social sciences, who were able to produce within 3 years at least 2 publications, which were recorded in Scopus*. According to the person ID of Scopus, researchers were selected who had published mainly according to the field classification of Scopus (ASJC) in the subfields of economics, psychology, sociology, and educational sciences (social sciences). The first publication should have been published before 2014 and publications should be available within 3 years. According to these criteria, 1,071 researchers from 12 universities and 4 subfields were selected. The combination of universities × subfields (12 × 4 = 48 and 47, respectively, since one combination was not available) was used as clusters in the multi-level model. Of the 1,071 social scientists, 291 (27.2%) were psychologists, 156 (14.6%) sociologists, 497 (46.1%) came from the economy and 127 (11.9%) from education. The academic age as the difference between the final year 2015 (time interval of the data) and the year of the first publication was on average 9.7 years (SD = 4.7) (Table 1).

The following bibliometric indicators served as items in the model: Number of top 5% publications, number of first author publications, citation of the highest cited paper, mean normalized citation score (number of papers with citation above the mean level of citations of a field), total citations for a 5-year window. A three-level statistical model was used (level 1: researcher, level 2: university × subfield, level 3: university).

Results

Sample description

The group of social scientists with Swiss university affiliations published 8.8 publications on average per capita during the study period, a minimum of 2 and a maximum of 104 (Table 2). With regard to the citation impact, 0.84 publications were on average in the top 5% percentile, about 4 publications were first author publications, about 3.4 publications were above the average of the total citations of a field. The citation of the most cited work amounted on average to 50 citations per capita. The h-index was 4 with a active research time of 7 years on average.

Table 1. Descriptive statistics per capita (N = 1,071 scientists)

Variable	Label	N	Mean	SD	Min	Mdn	P95%	Max
ITEM 1	Number of top5% publications	1,071	0.84	1.71	0	0	4	16
ITEM 2	Number of first author publications	1,071	4.05	4.50	0	3	12	42
ITEM 3	Citation of highest cited paper	1,071	49.97	108.53	1	21	169	1,639
ITEM 4	Mean normalized citation score	1,071	3.36	4.44	0	2	11	3,090
ITEM 5	Total citations (5-year window)	1,071	93.83	183.63	1	39	389	19
NPUB	Number of publications	1,071	8.82	8.94	2	6	27	104
AGE	Academic age	1,071	9.67	4.67	3	9	19	19
RTIME	Active research time	1,071	6.95	3.99	3	6	16	25
h	h-index	1,071	3.95	3.33	1	6	11	25

Note. SD = standard deviation, Min = minimum, Mdn = median, P95% = 95% percentile, Max = maximum.

With the exception of Item 2 (first authorship), psychologists had the highest mean values in all subfields. However, there were no significant differences between the subfields in the active research time.

Model comparison and model assumptions

In the first step, a model comparison was carried out to determine the model that best fitted the data (Table 2), once under the assumption of a Poisson distribution, once under the assumption of a Negative Binomial distribution.

As starting model a very restrictive one (M_1) was chosen, which assumed that all items had the same item difficulty β and item discrimination α. The restrictions were successively abandoned. The Deviance Information Criterion (DIC) served as the criterion for model comparison. The smaller the DIC, the better the model fits. In this respect, the best model was M_4, which assumes that the items have different item difficulties and discriminations. Models with a Negative Binomial distribution were clearly favored toward models with a Poisson distribution.

Of the *additional models*, a two-dimensional model (M_5) outperformed both, a model that allowed differences between subfields in the mean value of the person parameter and in the

item difficulties (M_6), and a model (M_7) that took into account the hierarchical structure of the data (3-level model). However, an additional multilevel model (M_8), in which the g-parameters were fixed in advance (informative prior), and which showed quite better convergence in the estimation process ("stationarity of Markov chains"), outperformed all other models and was selected as the final model. Eventually, a measurement model was obtained with a one-dimensional scale, where mean differences between subfields could be neglected.

Table 2. Model comparison with the Deviance Information Criterion (DIC).

	Factors				DIC	
MNo	Dimen.	Item difficulty β	Item discrimination α	Scale	Poisson	Negative Binomial
1	one	Equal	equal (=1)	equal	175,910.85	44,155.0
2	one	Unequal	equal (=1)	unequal	-	25,159.9
3	one	unequal	equal	unequal	42,132.3	25,051.6
4	one	unequal	unequal	unequal	41,952.2	**24,965.7**
Additional models						
5	two	unequal	unequal	unequal	26,124.9	24,850.4
6	M_4 + subfield differences in mean & difficulty				41,426.0	24,872.0
7	M_4 + multilevel				42,050.9	25,021.9
8	Final model: M_7 + g-components fixed				-	**24,631.4**

Note. "Equal" means that the respective item parameter value is constant cross items. "Unequal" means that the items vary in the respective item parameter. The lowest DIC values are bold faced.

Apart from the one-dimensionality, local stochastic independence is another prerequisite of the Rasch model. If the inter-correlations among the items are statistically controlled for the person parameters, the residual correlations should disappear (~0). In fact, the correlations between the items largely almost disappear, if one goes from the observed data (below diagonal) to the residuals (above diagonal) (Table 3). Thus, the assumption of local stochastic independence was widely confirmed.

The reliability of the scale amounted to r_{tt} = .84 and was rather high, but not perfect.

Table 3. Item inter-correlations (Spearman) for observed values (below diagonal) and for model residuals (above diagonal)

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	1.00	-.01	-.02	.11	.01
Item 2	.35	1.00	-.02	.12	-.00
Item 3	.51	.17	1.00	-.02	.03
Item 4	.78	.54	.39	1.00	.03
Item 5	.79	.40	.71	.79	1.00

Model interpretation

Instead of interpreting the model parameters, two figures are chosen in order to represent the model results. As explained above, the Poisson Rasch model consists of a binary core model and a frame model applicable to count data.

In the *binary core model* (Fig. 2) the probability to score excellently and to reach the maximum value of an item (e.g., the highest possible annual citation) is related to the person parameter, i.e. the researcher's performance capacity. With increasing person parameter value, the probability to score excellently increased. The turning points of the ICC, which are linked with vertical lines (Fig. 2), indicate the item difficulties.

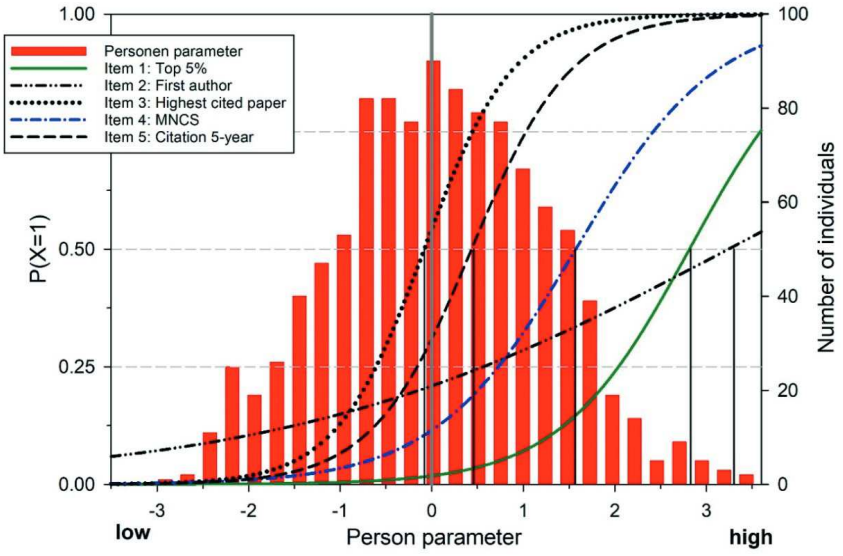

Figure 2. Item characteristic curve plot for the binary core Rasch model.

The following results can be formulated:

- *Person parameter*: Like most psychological characteristics (e.g., extroversion), the person parameters were symmetrically and normally distributed. In contrast, the bibliometric raw data are, actually, skewed distributed (e.g., Mutz & Daniel, 2012). For about 50% of the sample the person parameters were below 0 with probabilities less than 0.5 in all items. This means that half of the sample reached only half of the maximum annual rates in all items (e.g., highest citation).
- *Item parameter*: The two items for the raw citations (Item 3 and 5) showed the lowest item difficulties and the highest item discriminations of all items. It was easier for the social scientists to publish excellently in comparison to their colleagues from Switzerland (highest cited paper, citation 5-year window) than to publish excellently in international comparison regarding their field (top 5%, MNCS). The non-field normalized items distinguished better between researchers with high and low performance capacity than the field-normalized ones. Of low importance was the first authorship (Item 2), which showed both a low power to separate between researchers with high and low competency (item discrimination) and a high item difficulty. The items were more suitable for distinguishing scientists, which scored above average, than scientists, which scored below average.

The *frame model* allows the interpretation of the parameters in units of the items e.g., number of publications or citations (Fig. 3). In the ICC plot the annual accounts are related to the bibliometric quotient (BQ), which results from a simple linear transformation of the person parameters (Fig. 2) with mean value 100 and standard deviation of 15, which allows an formal not content-related interpretation of the BQ similar to an intelligence quotient (Fig. 3).

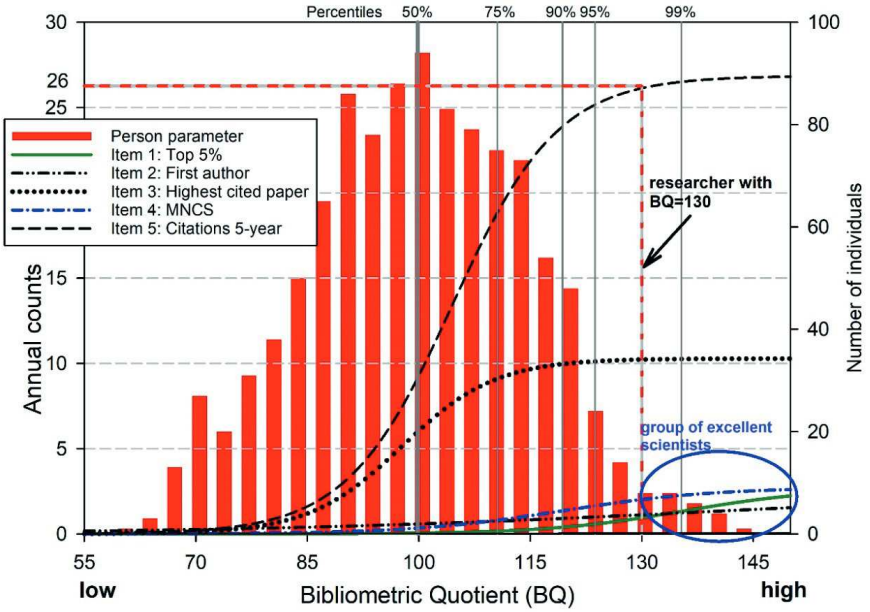

Figure 3. Item characteristic plot for the Poisson Rasch model for count data. Example: A researcher with BQ of 130 is likely to get 26 annual citations for his or her work.

The shrinkage correction of the Negative Binomial distribution and the "active research time" are not considered in the figure to facilitate the model interpretation.

According to Fig. 3 the BQ ranged from 59 (minimum) to 145 (maximum). A distinguishable group of very excellent scientists became visible, who had a BQ of over 130 (2 SD) and scored highly in all items. About 1% of the scientists had a BQ of 135 and higher.

Multilevel model

The final M_8 model also takes into account the fact that social scientists belonged to different subfields of social sciences (e.g., psychology) and different Swiss universities. The 2-level intra-class correlation (researcher, cluster) amounted to $\rho = .33$, i.e., 33% of the variability of the BQ was due to differences *between* the clusters subfields × universities (Level 2), and 67% to the variance *within* the clusters (Level 1). Only 7% of the total variance of the BQ was due to differences between universities (Level 3). The ranking of Swiss universities (Fig. 4) showed that the École Polytechnique Fédérale de Lausanne (EPFL) ranked first in the field of social sciences. However, the Goldstein-adjusted 95% credible intervals (Hox, 2010, p. 25) overlapped to such an extent that the differences in ranks could not be interpreted anymore.

Table 4. Correlations.

Correlation coefficient	Value
Intra class correlation ρ (researcher, cluster)	.33
Correlation (Spearman rank) of BQ with	
h-index	.72
total citations	.79
number of publications	.48

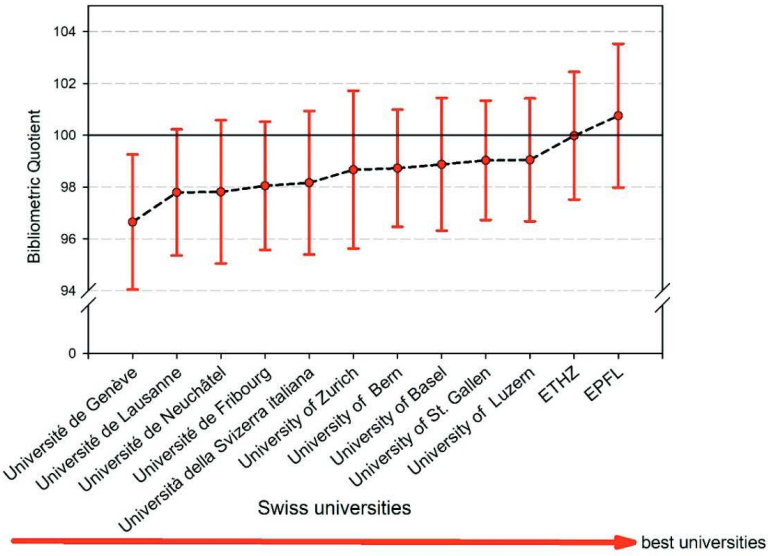

Figure 4. Ranking of Swiss universities from left to right (best universities) with Goldstein-adjusted 95% credible intervals.

Last but not least, classic bibliometric indicators such as the h-index or the total citations (Table 4) were highly correlated with the BQ (>.70).

Discussion

The amount of research articles, on how individual research performance can be measured, has increased significantly in recent years. Wildgaard et al. (2014) listed alone 108 author-level bibliometric indicators in their review. While in economy, sociology and information science the indicator approach is very common, in psychology and educational sciences scales are favored, which have to meet test theoretical requirements and have to take into account the fact that every measurement might be affected by measurement errors. It is assumed that the "researcher`s performance capacity" as trait and theoretical construct is responsible for the research output of a researcher. This can be hold for the indicator approach as well. With decisions such as the award of scholarships, it is not of primary interest which h-index a researcher has or how many top 5% articles he or she has published. Rather, it is a question of whether a researcher is able to influence his or her scientific field with his or her publications and to what extent the h-index or any other indicator or scale can say something about this competency.

The present contribution attempt to create a measurement scale on the base of test-theoretical concepts of psychology and educational sciences, which takes into account the multilevel structure of data (within and between institutions and subfields) with the following results:

- According to the Rasch model, the items formed a one-dimensional scale for assessing the "researcher's performance capacity" with a high, but not perfect reliability of $r_{tt} = .84$. The items were affected by measurement errors.
- Unlike bibliometric raw data the “researcher`s performance capacity” measured by the BQ was like other psychological characteristics approximately normally distributed. There was a group of very excellent social scientists with a BQ of over 130, who performed very well in all items.
- While it was easier for social scientists from Switzerland to perform well in comparison to their Swiss colleagues (raw citations), it was harder to perform well in the international arena (field-normalized citations). The items were most suitable for scientists, who scored above the average of the sample.
- Although around 33% of the variance was due to differences between the clusters subfield × university (67% within clusters), only 7% of the overall variance was actually due to differences between Swiss universities. A ranking in social sciences does not make any sense.
- The BQ is strongly related to classic bibliometric indicators, and it is not an artifact.

The results are limited, among other things, in that only a certain time interval could be used to estimate the BQ of a researcher. The results cannot necessarily be generalized to other countries. While the indicator approach at the level of institutions and countries has proved its worth, the question is whether alternative approaches at the individual level are needed that consider measurement errors. The model-oriented approach has the advantage of empirically testing certain questions of fairness, reliability, validity and invariance as empirical assumptions. A further question might be the influence of the sample selection process (only experienced researchers were focused) on the empirical results.

References

Abramo, G., Cicero, T., & D`Angelo, C. A. (2012). The dispersion of research performance within and between universities as a potential indicator of the competitive intensity in higher education systems. *Journal of Informetrics, 6*(2), 155-168. doi:https://doi.org/10.1016/j.joi.2011.11.007

Abramo, G., Cicero, T., & D`Angelo, C. A. (2013). Individual research performance: A proposal for comparing apples to oranges. *Journal of Informetrics, 7*(2), 528–539. doi:http://dx.doi.org/10.1016/j.joi.2013.01.013

Abramo, G., D`Angelo, C. A., & Grilli, L. (2015). Funnel plots for visualizing uncertainty in the research performance of institutions. *Journal of Informetrics, 9*(4), 954 -961 doi:https://doi.org/10.1016/j.joi.2015.08.006

Abramo, G., & D’Angelo, C. A. (2014). How do you define and measure research productivity? *Scientometrics, 2*(9), 1129-1144. doi:http://dx.doi.org/10.1007/s11192-014-1269-8

Alvarez, P., & Pulgarin, A. (1996a). Application of the Rasch model to measuring the impact of scientific journals. *Publishing Research Quarterly, 12*(4), 57-64. doi:https://doi.org/10.1007/BF02680575

Alvarez, P., & Pulgarin, A. (1996b). The Rasch model. Measuring information from keywords: The diabetes field. *Journal of the American Society for Information Science and Technology, 47*(6), 468-476. doi:https://doi.org/10.1002/(SICI)1097-4571(199606)47:6<468::AID-ASI7>3.0.CO;2-T

Alvarez, P., & Pulgarin, A. (1996c). The Rasch model. Measuring the impact of scientific journals: Analytical chemistry. *Journal of the American Society for Information Science and Technology, 47*(6), 458-467. doi:https://doi.org/10.1002/(SICI)1097-4571(199606)47:6<458::AID-ASI6>3.0.CO;2-U

Andrich, D. (2010). Rasch models. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed., Vol. 1, pp. 111-122). Oxford: Elsevier Science.

Bonaccorsi, A., & Cicero, T. (2016). Distributed or concentrated research excellence? Evidence from a large-scale research assessment exercise. *Journal of the Association for Information Science and Technology, 67*(12), 2976-2992. doi:https://doi.org/10.1002/asi.23539

Bornmann, L., & Marx, W. (2013). Evaluating individual researchers' performance. *European Science Editing, 39*(2), 39-40.

Bornmann, L., & Marx, W. (2014). Distributions instead of single numbers: Percentiles and beam plots for the assessment of single researchers. *Journal of the Association of Information Science and Technology, 65*(1), 206-208. doi:https://doi.org/10.1002/asi.22996

Bornmann, L., & Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: The avoidance of citation (ratio) averages in field-normalization. *Journal of Informetrics, 5*(1), 228-230. doi:https://doi.org/10.1016/j.joi.2010.10.009

Doebler, A., Doebler, P., & Holling, H. (2014). A latent ability model for count data and application to processing speed. *Applied Psychological Measurement, 38*(8), 587-598. doi:https://doi.org/10.1177/0146621614543513

Fox, J.-P. (2010). *Bayesian Item Response Modeling*. New York: Springer.

Harnad, S. (2008). Validating research performance metrics against peer rankings. *Ethics in Science and Environmental Politics, 8*(1), 103-107. doi:https://doi.org/10.3354/esep00088

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications* (2nd ed.). New York: Routledge.

Karlsson, A., Hammarfelt, B., Steinhauer, H. J., Falkman, G., Olson, N., Nelhans, G., & Nolin, J. (2015). Modeling uncertainty in bibliometrics and information retrieval: An information fusion approach. *Scientometrics, 102*(3), 2255-2274. doi: https://doi.org/10.1007/s11192-014-1481-6

Mutz, R., & Daniel, H.-D. (2012). Skewed citation distributions and bias factors: Solutions to two core problems with the journal impact factor. *Journal of Informetrics, 6*(2), 169-176. doi:https://doi.org/10.1016/j.joi.2011.12.006

Mutz, R., & Daniel, H.-D. (2018). The bibliometric quotient (BQ), or how to measure a researcher's performance capacity: A Bayesian Poisson Rasch model. *Journal of Informetrics, 12*(4), 1282-1295. doi:https://doi.org/10.1016/j.joi.2018.10.006

Mutz, R., & Wolbring, T. (2017). The effect of the "very important paper" (VIP) designation in Angewandte Chemie International Edition on citation impact: A propensity score matching analysis. *Journal of the Association for Information Science and Technology, 68*(9), 2139-2153. doi:https://doi.org/10.1002/asi.23701

Stone, C. A., & Zhu, X. (2015). *Bayesian Analysis of Item Response Theory Models using SAS*. Cary, NC: SAS Institute Inc.

Wechsler, D. (1958). *Measurement and Appraisal of Adult intelligence* (4th ed.). Baltimore: Williams & Wilkinds.

Wildgaard, L., Schneider, J. W., & Larsen, B. (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics, 101*(1), 125-158. doi:https://doi.org/10.1007/s11192-014-1423-3